

Sub A3

**SYSTEM AND METHOD FOR
ANALYSIS AND CLUSTERING OF
DOCUMENTS FOR SEARCH ENGINE**

ABSTRACT OF THE DISCLOSURE

5

A system and method for searching documents in a data source and more particularly, to a system and method for analyzing and clustering of documents for a search engine. The system and method includes analyzing and processing documents to secure the infrastructure and standards for optimal document processing. By incorporating

10 Computational Intelligence (CI) and statistical methods, the document information is analyzed and clustered using novel techniques for knowledge extraction. A comprehensive dictionary is built based on the keywords identified by the these techniques from the entire text of the document. The text is parsed for keywords or the number of its occurrences and the context in which the word appears in the documents. The whole document is identified by the knowledge that is represented in its contents. Based on such knowledge extracted from all the documents, the documents are clustered into meaningful groups in a catalog tree. The results of document analysis and clustering information are stored in a database.

20